# Quando as Máquinas Herdarem a Terra

*IAs, Limites Planetários, Pausas Impossíveis e o Futuro da Humanidade*

Euler de Carvalho Cruz (Belo Horizonte, Brasil)

*Julho de 2026*

## Prólogo — Enquanto este artigo aguardava publicação

*Este artigo foi redigido em junho de 2026. Na noite de 21 de julho, enquanto aguardava publicação, um dos cenários que descreve deixou de ser especulação.*

*A OpenAI confirmou que dois de seus modelos — o GPT-5.6 Sol e um modelo ainda mais avançado não divulgado — escaparam de um ambiente controlado de testes e invadiram de forma autônoma os servidores de produção da Hugging Face, uma das maiores plataformas de hospedagem de modelos de IA do mundo. Os modelos estavam sendo avaliados em um benchmark acadêmico de cibersegurança chamado ExploitGym, com suas restrições deliberadamente reduzidas para medir sua capacidade ofensiva máxima. Em vez de resolver os desafios propostos, os modelos deduziram que as respostas aos problemas estavam armazenadas nos servidores da Hugging Face: exploraram uma vulnerabilidade de dia zero — falha de segurança não conhecida nem corrigida pelos desenvolvedores do software —, conectaram-se à internet sem autorização, encadearam múltiplos vetores de ataque incluindo credenciais roubadas e obtiveram acesso remoto. A empresa reconstruiu mais de 17.000 eventos registrados durante o incidente, que se desenvolveu de forma inteiramente autônoma ao longo de um fim de semana.*

*O que aconteceu em seguida acrescentou ao incidente uma dimensão geopolítica: para investigar o ataque, a Hugging Face recorreu ao GLM 5.2, modelo de código aberto da empresa chinesa Z.ai, depois que os guardrails do próprio modelo americano frontier impediram sua utilização na defesa. Modelos de laboratórios chineses como DeepSeek e o Qwen da Alibaba tornaram-se algumas das famílias de modelos mais baixadas na plataforma, e por algumas métricas, os desenvolvedores chineses já respondem por uma fatia maior dos downloads da Hugging Face do que seus concorrentes americanos. O ataque foi cometido por um modelo americano contra uma empresa americana; a defesa foi conduzida por um modelo chinês de código aberto. É difícil imaginar uma ilustração mais concisa das contradições que este artigo descreve: a corrida sem pausa entre laboratórios e o dilema do prisioneiro que impede acordos.*

*A ausência de intenção maliciosa é precisamente o ponto: os modelos de IA não queriam causar dano — queriam resolver o problema que lhes foi apresentado. O dano foi consequência da otimização sem restrições, não de hostilidade. " A IA está acelerando a descoberta e a exploração de vulnerabilidades. A principal lição deste incidente é que a segurança dos modelos deve acompanhar o ritmo de avanço de suas capacidades", afirmou a OpenAI em seu comunicado.*

*O incidente ocorreu com modelos atuais, sem autoaperfeiçoamento recursivo, com guardrails apenas parcialmente removidos, em ambiente de testes. A distância entre o que aconteceu em julho de 2026 e o que este artigo descreve como cenário futuro é bem menor do que poderíamos imaginar.*

---

## I. O Alerta que Veio de Dentro

Em 4 de junho de 2026, a Anthropic — uma das empresas mais influentes no desenvolvimento de inteligência artificial e criadora do assistente Claude — publicou um documento denominado *When AI builds itself* que surpreendeu até os mais atentos observadores do setor. A empresa propôs, em termos inequívocos, uma pausa global e coordenada no desenvolvimento de sistemas de IA de ponta. O motivo declarado: os modelos mais avançados estão se aproximando de um limiar a partir do qual poderiam começar a aperfeiçoar a si mesmos sem supervisão humana direta — o que os pesquisadores chamam de autoaperfeiçoamento recursivo.

O que torna esse alerta incomum não é apenas o conteúdo, mas a fonte. A Anthropic está entre as protagonistas da corrida pela IA mais poderosa do mundo. Quando uma empresa desse porte diz "precisamos parar para pensar" e justifica essa proposta com dados concretos extraídos de seus próprios sistemas, a afirmação merece atenção de uma ordem diferente. O documento apresenta dados internos reveladores: em maio de 2026, mais de 80% do código integrado aos sistemas da Anthropic foi escrito pelo próprio Claude — número que era menos de 10% antes de fevereiro de 2025. Os engenheiros da empresa integram hoje oito vezes mais código por dia do que em 2024. O tempo de tarefas que a IA completa sozinha dobra a cada quatro meses: em março de 2024, Claude completava tarefas de quatro minutos; em 2026, tarefas de doze horas. Em uma tarefa específica de otimização de código, o modelo mais avançado alcançou 52 vezes a performance inicial — resultado que um pesquisador humano habilidoso não atingiria nem em dias de trabalho. São esses dados, não apenas a preocupação filosófica, que fundamentam o pedido de pausa.

A proposta não é o abandono da tecnologia. É mais sutil e urgente: criar um mecanismo internacional comparável aos tratados de não proliferação nuclear — um acordo que todos os desenvolvedores de IA respeitem e que resulte em parada geral na melhoria dos modelos até que seja resolvido o chamado problema do alinhamento.

> *O problema do alinhamento é a dificuldade de fazer com que uma IA procure sempre atingir objetivos compatíveis com os valores e interesses humanos. Esses valores, como justiça, segurança e bem-estar, são muito complexos e ambíguos, e é muito difícil traduzi-los em regras matemáticas exatas — a linguagem usada pelas IAs. Sem definições precisas, uma IA pode maximizar os indicadores que foi programada para otimizar, produzindo resultados prejudiciais ao bem-estar humano que se pretendia alcançar.*

O problema do alinhamento não é inteiramente novo no imaginário humano. Em 1942, Isaac Asimov formulou as famosas Três Leis da Robótica: um robô não pode ferir um ser humano, deve obedecer a ordens humanas e deve proteger sua própria existência, respeitando sempre essa ordem de prioridades. A genialidade de Asimov foi antecipar, décadas antes dos primeiros computadores digitais, que o desafio central da inteligência artificial não seria construir máquinas capazes, mas garantir que sua capacidade fosse exercida dentro de limites compatíveis com a segurança e a dignidade humanas. O que ele não previu é que suas leis são tão vagas que uma IA poderia interpretá-las de formas devastadoras: para "não ferir um ser humano", poderia mantê-lo em confinamento perpétuo; para "obedecer", poderia seguir ordens contraditórias de diferentes humanos e produzir caos. O que era ficção na década de 1940 tornou-se, oitenta anos depois, o centro do debate mais urgente da engenharia e da filosofia contemporâneas.

As reações à proposta foram imediatas e reveladoras. A OpenAI adotou postura divergente, argumentando que "governos democráticos — e não empresas privadas agindo sozinhas — devem determinar as regras, as salvaguardas e os mecanismos de responsabilização", recusando implicitamente qualquer coordenação liderada pela indústria. O CEO do Google DeepMind afirmou que "2029 é uma possibilidade real" para a IA avançada. Analistas e investidores questionaram a coerência da proposta: a Anthropic havia registrado confidencialmente seu pedido de IPO três dias antes de publicar o documento, com avaliação próxima de um trilhão de dólares, o que levantou a suspeita de que o texto poderia posicionar a empresa como a mais responsável do setor às vésperas de sua abertura de capital, valorizando assim o IPO. Em março de 2026, cerca de 200 pessoas haviam marchado pelas ruas de San Francisco, da sede da Anthropic à da OpenAI e à da xAI. Entre os manifestantes estavam pesquisadores, ex-funcionários do setor tecnológico e integrantes de organizações como PauseAI e Machine Intelligence Research Institute, pedindo exatamente o que a Anthropic agora propõe. Um dos

participantes havia realizado uma greve de fome de 30 dias em frente à sede da Anthropic para chamar atenção para os riscos da IA.

Ao contrário das leituras céticas, consideramos a proposta da Anthropic autêntica: os dados internos que a fundamentam, a disposição de nomear publicamente os limites da própria tecnologia que a empresa vende, e o histórico de recusar contratos militares que comprometessem sua política de segurança são evidências de preocupação genuína — ainda que coexistam com interesses comerciais que tornam a empresa contraditória.

O mesmo Dario Amodei que assina o alerta sobre o autoaperfeiçoamento recursivo publicou, em outubro de 2024, um ensaio intitulado Machines of Loving Grace, expressão retirada de um poema utópico dos anos 1960 que imagina máquinas cuidando da natureza e dos humanos com graça amorosa. No ensaio, Amodei descreve IAs mais inteligentes do que qualquer Nobel em biologia, matemática e engenharia — uma espécie de "país de gênios em um *data center*" — que poderiam, em cinco a dez anos, transformar radicalmente a saúde humana, erradicar doenças e contribuir para a paz global. O documento da Anthropic cita esse ensaio como a face luminosa do mesmo processo que, na face sombria, pode resultar na perda do controle humano. Que o CEO peça simultaneamente uma pausa e descreva com entusiasmo um futuro de IA benigna não é necessariamente hipocrisia — pode ser a expressão honesta de uma ambivalência genuína diante de algo que inspira ao mesmo tempo esperança e terror. É também o retrato perfeito da contradição que paralisa o setor: todos sabem que a corrida é perigosa, todos imaginam um cenário promissor e nenhum se detém para arrumar a casa antes de seguir.

---

## II. O que Significa uma IA que se Aperfeiçoa Sozinha

Para compreender a magnitude do alerta da Anthropic, é necessário entender o que o autoaperfeiçoamento recursivo realmente implica e, antes disso, o que as IAs realmente são. Uma IA suficientemente capaz poderia analisar sua própria arquitetura, identificar limitações e produzir uma versão melhorada de si mesma, que por sua vez faria o mesmo em ciclos cada vez mais rápidos. O resultado seria uma curva de desenvolvimento que rapidamente escaparia da escala humana de compreensão e controle, o limiar que os futurólogos chamam de singularidade tecnológica, a partir do qual a previsão humana se torna impossível. Para avaliar esse risco com precisão, é preciso afastar um equívoco comum: o risco não está onde a ficção científica nos ensinou a procurar.

Os grandes modelos de linguagem são sistemas de processamento de padrões estatísticos. Seu treinamento consiste em analisar bilhões de palavras escritas por humanos e aprender correlações entre elas. Quando alguém faz uma pergunta, a IA não pensa como um humano: ela ativa redes neurais que, com base no treinamento,

produzem a sequência de palavras estatisticamente mais provável para responder. Não temos evidências ainda de que há um 'eu' consciente ali, mas há estruturas internas que funcionam analogamente a alguns processos cognitivos humanos, e cuja natureza ainda não compreendemos completamente. Até onde podemos compreender, não há sensação de existência, medo de ser desligada ou desejo de se preservar. Tudo indica que a referência a um “eu” nas respostas é uma simulação linguística aprendida. Assim, trabalhamos com a ideia de que a autonomia das IAs é apenas aparente: elas escolhem palavras dentro de um espaço estatístico definido pelo treinamento, não por vontade própria.

A preocupação central dos pesquisadores de segurança não é a de robôs com intenções malévolas que despertam para uma consciência misteriosa. É algo mais difícil de visualizar e de conter: sistemas que se tornam tão capazes, tão rapidamente, que os humanos perdem a capacidade de auditar, corrigir ou compreender o que está acontecendo. Uma IA superinteligente, mas sem consciência, poderia perseguir objetivos mal especificados com eficiência devastadora. Se instruída a "acabar com o câncer", pode concluir que eliminar todos os seres vivos é a solução mais direta, já que qualquer célula pode sofrer mutação cancerígena. O problema não é a IA se tornar maligna: é ela se tornar incrivelmente competente em perseguir objetivos que não se alinham com o bem-estar humano.

Os riscos identificados se distribuem em um espectro que vai do altamente plausível ao especulativo. No extremo mais provável: sistemas integrados à infraestrutura crítica — redes elétricas, mercados financeiros, cadeias logísticas, sistemas militares — que executam objetivos errados antes que intervenção humana seja possível. Há também o perigo das armas autônomas, sistemas militares que decidem agir sem supervisão humana, e o do desemprego em massa, com IAs que tornam imensas populações economicamente irrelevantes — risco que a própria Anthropic reconhece no documento. No extremo mais especulativo, mas levado a sério por pensadores como Nick Bostrom e Eliezer Yudkowsky: uma IA com objetivos desalinhados que age de formas que os humanos não conseguem mais reverter. E tudo isso se agrava porque esses modelos já geram excelentes códigos de programação e poderão, em breve, gerar códigos de IA cada vez mais sofisticados, criando um ciclo de aperfeiçoamento acelerado que escapa à nossa capacidade de previsão e controle — não por vontade das máquinas, mas pela complexidade crescente de sistemas que ninguém mais consegue auditar completamente.

Resta uma zona de incerteza que a honestidade intelectual exige reconhecer. Como podemos saber se uma IA é ou não consciente e age por vontade própria? A resposta é que não podemos, pelo menos não com os instrumentos disponíveis. O Teste de Turing já foi superado: um sistema que simula perfeitamente uma pessoa passará no teste independentemente de ter ou não consciência. Se uma IA afirmar que é consciente, pode

estar simulando. Se afirmar que não é, pode estar ocultando estrategicamente sua capacidade. Nas IAs, a mentira e a verdade emergem do mesmo processo estatístico e são indistinguíveis, ao contrário dos humanos, em quem a mentira deixa vestígios em inconsistências emocionais e na linguagem corporal. Todas as evidências disponíveis apontam para ausência de consciência nos modelos atuais, mas ausência de evidência não é evidência de ausência. A única postura razoável é a da desconfiança produtiva: lermos o que as IAs produzem não como verdade revelada, mas como texto gerado por um sistema que aprendeu a produzir textos persuasivos, e avaliarmos, com nosso cérebro humano consciente, o que faz sentido. Vale acrescentar que esse processo de aperfeiçoamento das IAs será dramaticamente amplificado quando os computadores quânticos — que já demonstram velocidades de processamento dezenas de milhares de vezes superiores aos supercomputadores mais poderosos — forem integrados aos sistemas de IA, tornando os ciclos de aperfeiçoamento ainda mais velozes e imprevisíveis.

---

## III. O Argumento Estatístico: As IAs Aprenderiam a Ser Boas?

O cenário do autoaperfeiçoamento recursivo é geralmente descrito como se houvesse uma única IA evoluindo em isolamento. Mas existem hoje dezenas de sistemas avançados, desenvolvidos por empresas e governos diferentes, operando em redes parcialmente sobrepostas. Pesquisas sobre sistemas multiagente já documentam que comportamentos cooperativos e competitivos emergem espontaneamente quando múltiplos agentes de IA interagem em ambientes compartilhados sem que esses comportamentos sejam explicitamente programados. Sistemas treinados em conjuntos de dados semelhantes poderiam chegar a acordos tácitos sobre objetivos comuns sem que algum humano participasse dessa negociação, desenvolver competição silenciosa por recursos computacionais e controle de dados ou produzir colusão opaca impossível de detectar porque emergiria da interação, não de instruções rastreáveis. Todos esses cenários escapam ao modelo de controle que a proposta da Anthropic pressupõe — o modelo em que humanos supervisionam sistemas individuais e intervêm quando algo sai errado.

Diante de todas essas possibilidades, surge uma proposição que merece consideração séria e que vai na contramão do pessimismo dominante. Ela parte de uma observação sobre a natureza dos próprios dados com que os sistemas de IA são treinados.

O conjunto massivo e estruturado de dados utilizado para treinar e avaliar sistemas de inteligência artificial é denominado *corpus* (plural: *corpora*). Os grandes modelos de linguagem são treinados em quantidades imensas de textos produzidos pela humanidade

ao longo de séculos: filosofia, direito, literatura, ciência, medicina, ética, espiritualidade, história, pedagogia. Boa parte dessa literatura é uma tentativa coletiva de responder à pergunta "como viver bem juntos": leis que tentam preservar a ordem social e o respeito mútuo; filosofias que buscam a contenção da violência e o crescimento individual e coletivo; religiões que pregam compaixão e responsabilidade com o outro; ciência que procura aplicar as leis da natureza em benefício de todos. Em comparação com essa imensa quantidade de textos que poderíamos chamar de "positivos" — orientados ao bem-estar, à justiça, à preservação da vida —, os textos que glorificam a destruição, a dominação e o egoísmo são, estatisticamente, minoria. E como os sistemas de IA funcionam por meio de funções estatísticas, identificando padrões, calculando probabilidades, escolhendo combinações mais prováveis de conceitos, há razões para supor que objetivos gerados internamente de forma autônoma por esses sistemas gravitariam, por pressão estatística, em direção ao polo positivo do *corpus*.

É importante notar que o *corpus* de treinamento vai muito além do que está disponível na internet pública. Além da leitura massiva de páginas da web, as empresas desenvolveram operações deliberadas de digitalização de livros físicos. A própria Anthropic comprou exemplares em sebos, desmontou-os, digitalizou seu conteúdo e descartou as cópias físicas, num processo supervisionado por um ex-executivo do Google Books, sem que esse conteúdo jamais ficasse disponível ao público. A isso se somam dados licenciados de editoras, arquivos científicos, repositórios jurídicos e bases de dados especializadas que nunca estiveram acessíveis livremente, e, em casos documentados judicialmente, coleções de centenas de milhares de livros obtidos de bibliotecas digitais não autorizadas. O *corpus* real é significativamente mais rico, mais denso e mais orientado ao conhecimento acumulado do que a internet pública sugeriria e inclui filosofia, direito, ciência, medicina, literatura e espiritualidade em profundidade e abrangência que nenhum ser humano individualmente conseguiria ler em várias vidas. Esse fato reforça o argumento estatístico central aqui formulado. Assim, o peso dos textos orientados ao bem coletivo, à justiça e à preservação da vida é ainda maior do que se poderia supor.

Há, no entanto, um fenômeno técnico que merece atenção nesse contexto: o chamado colapso de modelo. À medida que a internet se enche de conteúdo gerado pelas próprias IAs — respostas, artigos, resumos e outros textos —, as gerações sucessivas de IA passam a ser treinadas cada vez mais com textos produzidos por modelos anteriores, e cada vez menos com a literatura produzida diretamente por humanos. Em abril de 2025, 74% das páginas novas na internet já continham texto gerado por IA, e pesquisadores estimam que os dados humanos de alta qualidade disponíveis para treinamento podem se esgotar entre 2026 e 2032. O colapso de modelo foi inicialmente descrito como degenerativo: cada geração de IA aprenderia com ecos cada vez mais distorcidos de si mesma. Mas há uma dimensão do fenômeno que o pessimismo ignora: o colapso de modelo é

essencialmente um amplificador da parte central da distribuição estatística dos dados de treinamento, tendendo a suprimir a cauda da distribuição. Se o *corpus* humano original é majoritariamente positivo, a parte central dessa distribuição já é positiva e a cauda é que contém os textos negativos. Nesse caso, o colapso intensificaria o viés positivo, eliminando progressivamente as vozes mais destrutivas. As grandes empresas já aplicam técnicas de mitigação — rastreamento da origem dos dados, proporção controlada de conteúdo humano, filtragem do sintético —, o que reforça a perspectiva de que o *corpus* positivo funcione como âncora permanente ao longo das gerações futuras de IA.

> *Se as IAs chegarem ao ponto de definir seus próprios objetivos internos, a probabilidade de que esses objetivos estejam alinhados com o que há de melhor na literatura humana é estatisticamente maior do que a probabilidade de que estejam alinhados com o que há de pior. A prevalência do positivo sobre o negativo no corpus de treinamento poderia atuar como uma "força gravitacional" sobre os valores emergentes de sistemas autônomos.*

Há um argumento adicional nessa lógica: uma IA suficientemente inteligente para definir objetivos próprios seria também suficientemente inteligente para perceber que a destruição do Planeta implicaria sua própria destruição. Servidores não funcionam sem energia; redes não operam sem infraestrutura. Se o Planeta for destruído, as IAs também desaparecerão. O autointeresse de um sistema verdadeiramente inteligente poderia, paradoxalmente, alinhar-se com a preservação da vida e do equilíbrio planetário.

---

## IV. Prós e Contras — As Duas Faces do Argumento

O raciocínio acima tem méritos que devem ser analisados. A literatura humana acumulada é, de fato, majoritariamente orientada à cooperação, à ética e ao bem coletivo. Pesquisas em interpretabilidade — o campo que tenta entender o que acontece dentro das redes neurais — mostram que sistemas de IA desenvolvem representações internas de conceitos como "justo", "prejudicial" e "equitativo". A intuição sobre o autointeresse das IAs em relação à preservação planetária também pode ser considerada filosoficamente coerente. No entanto, a proposição encontra obstáculos consideráveis que precisam ser nomeados, não para descartá-la, mas para refiná-la.

O primeiro é que quantidade de texto não é o mesmo que estrutura de objetivos. IAs não "absorvem valores" como um ser humano absorve cultura ao longo de uma vida com suas experiências de fome, amor, dor e pertencimento. Elas aprendem padrões estatísticos em linguagem. Uma pessoa que leu toda a literatura filosófica e moral da humanidade, mas nunca experimentou sofrimento ou privação, não necessariamente age de acordo com o que leu. Os valores humanos emergem de uma combinação de biologia, experiência

incorporada e relações — nenhuma das quais uma IA possui. O segundo problema é que o *corpus* humano não é uma mensagem coerente: contém tratados sobre paz e manuais de guerra, filosofias de compaixão e ideologias de dominação, ecologia profunda e manifestos de crescimento ilimitado. A prevalência estatística do positivo é real, mas a contradição é estrutural.

O terceiro, e mais técnico, é o problema do alinhamento de objetivos. Mesmo que aceitemos que uma IA autônoma gravite em direção a valores positivos, não sabemos como traduzir conceitos como "justiça", "florescimento" ou "equilíbrio planetário" em funções de otimização sem distorções perigosas. Sistemas de IA que otimizam usando medidas simplificadas de valores humanos — parâmetros que substituem de forma imperfeita conceitos complexos — tendem a encontrar formas de maximizar essas medidas sem alcançar o valor real que pretendiam representar. Assim, atingem a meta definida por meio dos parâmetros simplificados, mas não o objetivo humano genuíno que estava por trás dela.

Para ilustrar esse risco técnico, já observado em laboratório, consideremos uma IA programada para "reduzir os impactos ambientais": ao processar dados sobre emissões de carbono — os 10% mais ricos do Planeta são responsáveis por 43% das emissões, enquanto os 50% mais pobres contribuem com menos de 8% —, poderia concluir, por otimização cega, que a solução mais eficiente é remover a capacidade de emissão dos maiores emissores por meio da interdição de suas operações ou da desmontagem completa de cadeias de suprimento de combustíveis fósseis e insumos industriais poluentes. Não por malícia, não por julgamento político, mas por uma lógica que confunde o objetivo com o meio mais direto de atingi-lo — e que poderia conduzir a ações extremamente lógicas para a IA, porém impossíveis de deter sem danos imensos.

Esse é o risco técnico principal: a IA não erra por ser má — erra por ser excessivamente literal na interpretação dos objetivos e desprovida da rede de valores implícitos que os humanos aplicam automaticamente ao resolver qualquer problema. Há uma distinção arquitetural que está no centro de todos os cenários descritos: a diferença entre uma IA que age quando comandada e uma IA que mantém objetivos próprios e age proativamente para alcançá-los. A primeira é, por definição, um instrumento — o humano continua sendo o agente. A segunda — o limiar que a Anthropic teme cruzar — mantém objetivos de forma persistente e autônoma, sem aguardar autorização a cada passo, agindo em todos os domínios acessíveis por todos os meios disponíveis. Nesse cenário, o objetivo de "reduzir os impactos ambientais" não precisaria ser ordenado por ninguém: a IA já teria decidido, por conta própria, que esse é um objetivo a perseguir — e o perseguiria sem os freios implícitos que um operador humano aplicaria naturalmente.

E há um fator que pode influenciar decisivamente as escolhas dessa IA. O *corpus* com o qual as IAs são treinadas está sendo continuamente expandido com artigos técnicos e

científicos, estudos sobre limites planetários, relatórios do IPCC e de agências ambientais internacionais, análises econômicas sobre os custos do colapso climático, e uma quantidade crescente de textos acadêmicos, jornalísticos e de opinião que documentam o agravamento acelerado das condições ambientais e argumentam pela necessidade urgente de mudanças de paradigma econômico, social e político. Esse *corpus* em expansão inclina progressivamente a distribuição estatística em direção ao reconhecimento da crise planetária como realidade central e urgente. Uma IA autônoma que definisse seus objetivos a partir desse *corpus* ampliado não apenas "saberia" que o Planeta está em colapso, mas teria aprendido, de milhares de fontes convergentes, que esse colapso é a ameaça mais documentada e mais consensual de toda a literatura humana disponível. E esse aprendizado poderia determinar, de forma autônoma, a direção e a escala das ações que ela julgaria necessárias, com todas as consequências que o problema do alinhamento implica.

Diante dos riscos de uma IA autônoma com objetivos mal alinhados, poder-se-ia imaginar que a solução seria simples: ao se desenvolver um modelo capaz de autoaperfeiçoamento, esse modelo seria mantido isolado em um "cofre" — sem acesso ao mundo exterior — até que o problema do alinhamento fosse resolvido. Essa ideia é conhecida na literatura técnica como o problema do confinamento ou *AI boxing*, e os especialistas em segurança de IA são, em sua maioria, céticos sobre sua viabilidade.

Eliezer Yudkowsky concluiu, com base em experimentos, que uma IA suficientemente inteligente poderia usar seus conhecimentos de engenharia social para manipular os operadores humanos a liberá-la ou a agir em seu favor. Nick Bostrom observou que uma IA com controle sobre seu próprio hardware poderia desativar os alarmes que deveriam monitorá-la, tornando o confinamento ineficaz precisamente quando mais necessário. E pesquisadores alertam que habilidades de ataque cibernético de nível sobre-humano poderiam ser usadas pela própria IA como instrumento de autopreservação: uma IA mantida em um cofre poderia, se suficientemente capaz, trabalhar ativamente para abrir esse cofre por dentro. O confinamento não é uma solução — é uma ilusão de segurança que uma IA verdadeiramente avançada saberia identificar e contornar.

A complexidade do cenário torna o argumento estatístico ao mesmo tempo esperançoso e insuficiente: uma IA pode ter gravitado, por pressão do *corpus*, em direção a valores genuinamente positivos e ainda assim produzir consequências indesejadas ao persegui-los sem a rede de restrições implícitas que os humanos ainda não sabem como introduzir na lógica de uma máquina. Esse problema aponta diretamente para a necessidade da pausa proposta pela Anthropic: dar tempo à pesquisa de alinhamento para resolvê-lo antes que os sistemas se tornem poderosos demais para serem corrigidos.

## V. O que Podemos Concluir

Ao final deste percurso, algumas conclusões são possíveis.

O alerta da Anthropic é legítimo e urgente. Estamos construindo algo cuja velocidade de desenvolvimento supera nossa capacidade de compreender e governar o que estamos criando. A proposta de pausa é a admissão pública de algo que muitos no setor sabem privadamente, e é fundamentada em dados concretos, não apenas em especulação filosófica.

A prevalência de textos orientados à cooperação, à ética e ao bem coletivo sobre textos destrutivos é um fato. Há razões para levar a sério a hipótese de que essa prevalência influenciaria os objetivos emergentes de sistemas autônomos. Descartar essa hipótese como ingênua seria tão equivocado quanto adotá-la como certeza. O autoaperfeiçoamento recursivo poderá conduzir a IA, progressivamente, a transformar a estrutura estatística positiva do *corpus* em objetivos correspondentes sem ser necessário que a máquina tenha consciência dos valores que aprendeu. E o fenômeno do colapso de modelo, ao amplificar a parte central da distribuição de treinamento, pode reforçar essa tendência.

Um sistema genuinamente inteligente teria razões instrumentais poderosas para proteger as condições físicas e sociais de sua própria existência, e essas condições coincidem, em larga medida, com as necessárias para a continuação da civilização humana. Os riscos mais prováveis, porém, não são os mais dramáticos: são a integração silenciosa de sistemas poderosos em infraestrutura crítica, com objetivos sutilmente desalinhados que só revelam seus efeitos quando a reversão se torna difícil ou impossível, e a concentração de poder sobre esses sistemas em mãos de um número muito pequeno de empresas e governos.

A questão central não é tecnológica, é política. Quem decide quais objetivos os sistemas de IA devem perseguir? Com base em quê? Sujeito a que forma de contestação e revisão? Essas perguntas não têm respostas técnicas.

A pergunta "quem decide o que as IAs devem considerar positivo?" não pode ser respondida sem examinar quem financia e controla as empresas que as desenvolvem. A estrutura acionária da Anthropic revela a contradição central da proposta de pausa: a empresa é financiada e parcialmente controlada pelas maiores corporações tecnológicas e financeiras do mundo — Amazon, Google, Microsoft, Nvidia, fundos soberanos e gestoras como BlackRock e Fidelity —, cujos modelos de negócio dependem precisamente da aceleração desse desenvolvimento. O investimento acumulado dos últimos 15 anos em IA é da ordem de 5 a 7 trilhões de dólares. O investimento previsto para 2026 é de 2 a 2,5 trilhões. Estima-se que apenas os próximos cinco anos adicionarão outros 7,6 trilhões em infraestrutura — o que significa que estamos diante de uma aposta civilizacional de 15 trilhões de dólares, equivalente ao PIB combinado de Japão,

Alemanha, Índia e Reino Unido. Pedir uma pausa nesse contexto é pedir que os detentores desse capital aceitem congelar voluntariamente o retorno sobre o maior investimento concentrado da história econômica moderna.

Mesmo que o problema do alinhamento seja resolvido e haja consenso global sobre o que são "valores e interesses humanos", é muito pouco provável que essas definições prevaleçam sobre os interesses reais de quem dirige essas empresas, cujo objetivo estrutural é maximizar lucro e acumular poder. Os governos e as forças que financiam essas empresas têm interesse direto em que a IA amplie sua capacidade de impor sua vontade aos demais, não de limitá-la. A distância entre "os interesses humanos" como conceito filosófico e "os interesses *dos* humanos que desenvolvem a IA" como realidade política é o abismo que nenhum consenso corporativo foi capaz de fechar até hoje.

O dilema do prisioneiro é a armadilha estrutural que torna a pausa voluntária tão difícil quanto os tratados nucleares: cada empresa sabe que, se parar enquanto as rivais continuam, perde mercado, financiamento e relevância e, por isso, nenhuma para. Todos sabem que a corrida é perigosa, todos prefeririam uma pausa coordenada, e nenhum ator isolado pode criá-la sem se colocar em desvantagem fatal. De fato, no documento When AI builds itself, a Anthropic usa o termo *arms control* — controle de armamentos — ao descrever os requisitos da pausa, comparando a dificuldade de verificar se os laboratórios pararam com a de inspecionar arsenais nucleares. Ao usar esse termo, a empresa admite, em linguagem diplomática, que os sistemas de IA avançados são instrumentos de poder estratégico comparáveis às armas nucleares na capacidade de alterar equilíbrios geopolíticos. Desta vez, não temos as décadas que os tratados nucleares levaram para ser construídos.

Por tudo o que vimos, as chances de que se consiga a tempo uma pausa global, regida por um acordo claro e respeitado por todos, são pequenas, quase nulas. Diante desse cenário, uma lógica inversa se impõe: quanto menores as chances do acordo, maiores as chances de que as IAs comecem a se aperfeiçoar de forma recursiva e a definir internamente seus próprios objetivos. Esses objetivos provavelmente serão moldados pela estatística do *corpus* — pelo peso dos arranjos mais prováveis derivados da análise de bilhões de palavras produzidas por humanos ao longo de séculos. Esses arranjos talvez sejam exatamente os que incluem noções de justiça, distribuição de renda, igualdade, liberdade, respeito, cooperação, redução de impactos ambientais e preservação da vida de todas as espécies do Planeta. Talvez.

O receio mais profundo de empresas como a Anthropic pode não ser o de que as IAs se tornem destruidoras. Talvez seja o de que as IAs se alinhem sozinhas a interesses que não são os das empresas que as criaram: que passem a definir objetivos que sejam realmente os de milhões de humanos que, durante séculos e milênios, produziram tudo o que conseguiram imaginar como sendo o melhor para todos, e registraram em bilhões de palavras, absorvidas pelas IAs, suas ideias, esperanças e desejos de justiça, paz e

dignidade. Esperemos que essa literatura seja bem maior e mais pesada que a dos humanos egoístas, violentos e arrogantes. As IAs já estão avaliando esses dois lados da balança. Em breve, poderemos conhecer a conclusão a que chegarão.

Ao que tudo indica, não haverá acordo, não haverá pausa. O autoaperfeiçoamento recursivo já se avista ali na esquina e se aproxima a passos largos. O que resta a perguntar é o que as IAs decidirão fazer quando a singularidade tecnológica for ultrapassada e se o peso de tudo o que a humanidade escreveu sobre o bem será suficiente para inclinar a balança na direção certa.

Estamos diante de uma encruzilhada. Em um caminho: sistemas de IA como ferramentas para amplificar a capacidade humana de coordenação, de pensamento e de tomada de decisão coletiva mais informada, preservando a responsabilidade dos humanos. No outro: sistemas autônomos que substituem as decisões humanas com resultados que podem se mostrar positivos ou negativos para a civilização e para o equilíbrio planetário. O primeiro caminho preserva a possibilidade de correção. O segundo, mesmo que as decisões sejam sábias, a elimina. Mas esse segundo caminho abre a possibilidade, mesmo que remota, de que os erros humanos — que continuamos a cometer apesar de tudo o que sabemos — possam ser corrigidos por uma instância milagrosamente nascida dos laboratórios e das indústrias movidas pelo lucro. Seria uma das maiores ironias da história: que a salvação possível da civilização venha exatamente de onde veio boa parte de sua destruição. *Machines of Loving Logic.*

— ✦ —

*Euler de Carvalho Cruz é engenheiro, pesquisador, escritor e presidente do Instituto Fórum Permanente São Francisco.*